Statistics and Application 统计学与应用, 2019, 8(2), 277-281
Published Online April 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.82031

Hans汉斯

Research on Stock Classification Based on K-Means Clustering

Menghan Zhang, Mu Yang, Zai Chen

Hangzhou Dianzi University, Hangzhou Zhejiang
Email: 757268102@qq.com, lfyangmu@126.com

Received: Mar. 19th, 2019; accepted: Apr. 3rd, 2019; published: Apr. 10th, 2019

Abstract

With the development of China's economic market and the gradual improvement of the stock market, there are about 2000 stocks in China's stock market at present, and more and more people regard stocks as a major investment method. We always expect to obtain the maximum benefit with the minimum risk when investing. Facing the huge stock market and complicated stock data, cluster analysis is particularly important for reasonable analysis and selection of stocks. In this paper, we use k-means clustering to cluster 20 randomly selected stocks, then we analyze various types and give corresponding investment suggestions.

Keywords

K-Means Cluster, Stock Analysis, Investment Recommendations

基于K-Means聚类的股票分类研究

张梦涵，杨　牧，陈　宰

杭州电子科技大学，浙江　杭州
Email: 757268102@qq.com, lfyangmu@126.com

收稿日期：2019年3月19日；录用日期：2019年4月3日；发布日期：2019年4月10日

摘　要

随着中国经济市场的发展以及股票市场的逐步完善，目前在中国股票市场上有约2000支股票，越来越多的人将股票作为一种主要的投资方式。人们在投资时总期望以最小的风险获取最大的利益，面对庞大的股票市场和繁杂的股票数据，要想对股票进行合理的分析和选择，聚类分析就显得尤为重要。在本文中，我们采用了K-means聚类法对随机选择的二十支个股进行了聚类，并对各类股票进行了分析，给出了相

文章引用: 张梦涵, 杨牧, 陈宰. 基于 K-Means 聚类的股票分类研究[J]. 统计学与应用, 2019, 8(2): 277-281.
DOI: 10.12677/sa.2019.82031

应的投资建议。

关键词

K-Means聚类，股票分析，投资建议

1. 前言

股票市场作为市场经济发展的重要部分，逐渐朝着规范化、成熟化的方向发展，在数据繁杂信息爆炸的时代背景下，股票市场会出现相应的不确定因素。股市涨跌无常、难以预测，想要在股市中以较低风险获取一定的回报，需要合理利用数据分析的方式来分析数据，保证能够从数据中提取有效的股票市场信息。

聚类是数据挖掘领域非常重要的一项技术，它可以发现数据很多潜在的信息和价值。针对股票市场，应用聚类分析模型来进行深入挖掘，可为投资者提供可靠的帮助。文献[1]通过运用聚类的方法对股票进行分析评价；文献[2]不仅运用聚类的方法对股票进行了分析，还根据聚类所得的数据信息对股票进行了投资分析；文献[3]运用 K-means 聚类单独对 ST 股票进行了分类研究，给出了相应的投资策略。文献[4]针对经典 K-means 聚类算法过于依赖初始聚类中心和易陷入局部最优的不足，提出一种带有学习能力的人工蜂群算法与 K-means 迭代相结合的聚类算法，使聚类收敛速度更快，稳定性更强，聚类精度也更高；文献[5]分析了聚类模型在股票市场应用过程中的优势与局限性。

我们利用 k-means 聚类方法对随机选取的 20 支股票进行了聚类，并给出了各类股票的特性，基此，给一般投资者提出投资建议。

2. 研究目的与方法

2.1. 研究目的

本文选取二十支个股作为研究对象，对它们进行聚类分析，从而对这些个股公司进行划分，并利用对每类股票的研究成果分析股票市场，给广大的投资者一个直观的理论参考以及一定的决策依据。

2.2. 研究方法

本文采取 K-means 聚类算法，以个股的收益率为研究数据，采用欧氏距离作为分类标准，将二十支个股分为三大类，并对每类进行图表分析，给出直观的分类结果。

3. 经典 K-means 聚类算法

把数据集 $\Omega=\{x_1,x_2,\cdots,x_n\}$ 划分为多个子集 $C_j\ (j=1,2,\cdots,m)$。其中 $x_i\ (i=1,2,\cdots,n)$ 有 m 个属性，m 即为聚类个数，同时满足以下约束条件：

$$\bigcup_{j=1}^{m} C_j = \Omega \tag{1}$$

$$C_j \neq \varnothing, j = 1,2,\cdots,m且i \neq j \tag{2}$$

$$C_i \cap C_j = \varnothing, i, j = 1,2,\cdots,m \tag{3}$$

基于划分的聚类质量评估函数：

$$J = \sum_{j=1}^{m}\sum_{i=1}^{n_j} d\left(x_i, z_j\right), x_i \in C_j \tag{4}$$

其中：n_j 是子类 C_j 中的样本总数；z_j 表示子类 C_j 的聚心，是该子类所含样本的均值 $z_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i$；$d\left(x_i, z_j\right)$ 表示样本 x_i 与所属子类 C_j 的中心 z_j 的距离；J 表示全部类内距离和，J 的值越小，说明聚类效果越好。

经典 K-means 聚类算法描述如下：

输入：含有 n 个样本的数据集 Ω，聚类个数 m。

输出：满足聚类质量评估函数值最小的 m 个聚类。

I：从数据集 Ω 中随机选择 m 个样本作为初始聚心；

II：根据类 $C_j\left(j = 1,2,\cdots,m\right)$ 中所含样本的均值，计算每个样本到各类聚心 z_j 的距离，把其归到距离最小的类；

III：重新计算类 $C_j\left(j = 1,2,\cdots,m\right)$ 的聚心 z_j；

IV：计算本次迭代的聚类质量评估函数 $J(t)$ 并和上次迭代的聚类质量评估函数 $J(t-1)$ 比较；若两者之差满足阈值则算法结束，否则转到 II 继续执行。

4. 数据来源

数据主要来源是通过同花顺财经网站提供的个股指标信息。同花顺财经网站是国家银监会认证的互联网企业，主要是股市信息与交易平台，数据具有专业性、及时性、可靠性、多样性等特点。我们在同花顺财经网站的个股市场中随机选取了二十支股票作为研究样本，收集了其 2017 年 3 月~8 月的收盘价用于聚类研究。我们将数据进行了处理，以三天为一个周期，用第三天的收盘价减去第一天的收盘价的差值除以第一的收盘价得到该周期的收益率，从而获得二十支股票半年内的收益率数据。

5. 聚类结果分析

利用 MATLAB 程序编辑窗口，利用聚类分析中的经典 K-means 的聚类方法，采用欧式距离为度量方式，对收集到的 20 支个股数据进行聚类分析，将其可以分为三类。分类结果如表 1 所示。

通过观察可以得到，第一类股票三维丝，前期它的走势趋于平稳，波动幅度很小，但在后期它呈增

Table 1. Classification results of K-means clustering
表 1. K-means 聚类法分类结果

类别	股票名称	股票总数
第一类	三维丝	1
第二类	中钢国际	1
第三类	凤凰光学、龙溪股份、中葡股份、飞凯材料、八一钢铁、北方矿业、安达科技、美丽生态、紫金矿业、中国核电、中国银行、英力特、好利来、建新股份、韶钢松山、顺丰控股、恒通科技、三特索道	18

长状态，且上升幅度非常大，我们做出收益率的折线图更直观的反映股票波动，如图 1 所示。

对于三维丝这类增长型股票，其增长有提高资本的配置效率的作用，并能进一步促进社会经济的增长。资本配置效率的提高意味着在社会资本总量不发生改变的情况下，货币资本能够在长期利润信号驱动下在股票市场中的各个产业部门、企业与个人之间高效流动[6]。对个人来说，增长型股票是高效获利的最佳捷径；对企业来说，自身股票的增长对于公司潜在价值和融资有着良好的导向；而对各个产业部门，资金配置效益好、效率高能有利推动经济增长的集约化。

Figure 1. Line chart: stock yield of Savings
图 1. 三维丝股票收益率波动折线图

第二类股票中钢国际，它的整体波动幅度较大，且存在一天的收益率波动幅度超过了 0.3，波动折线图如图 2 所示。

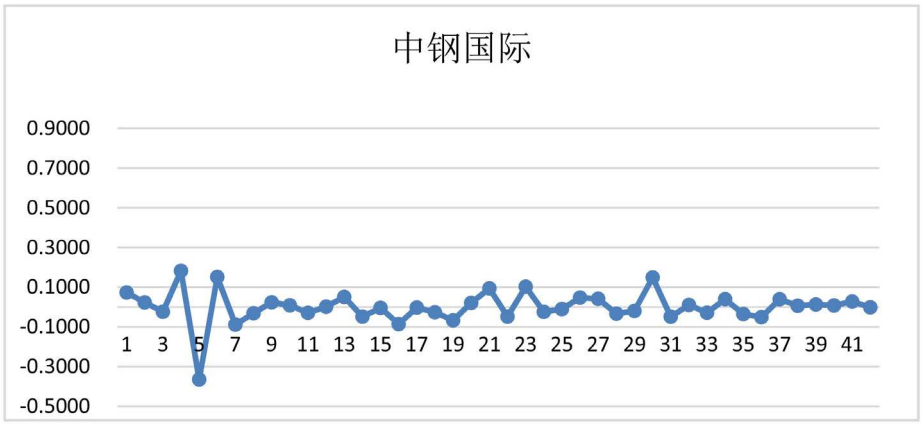

Figure 2. Line chart: stock yield of Sinosteel
图 2. 中钢国际股票收益率波动折线图

第二类股票出现了明显的股价下跌，说明该类上市公司可能存在盈利能力差、资产质量差的缺陷，没有发展前景，而且稳定创造现金能力差。投资者在操作时要尽量谨慎，建议最好规避风险。

第三类股票，它们的整体波动幅度较小，波动范围都在 0.3 范围以内。波动折线图如图 3 所示。

对第三类股票，狭义上来讲，股票价格的稳定象征着股票市场的稳定，从本质上来讲，宏观经济运行状态和宏观政策是影响保持股票市场稳定的根本因素。今年来，行为金融学的发展，投资者的情绪越

来越受到关注，投资者的情绪会极大的影响其决策进而影响资产价格。而影响投资者的情绪的主要来源是国家各类决策的投放因此，稳定型股票也不妨是一种检验国家投放决策正确与否的方法。

综上所述，投资者对于处理过的三类股票要进行区别对待，具体问题具体分析，投资者根据自己的风险厌恶偏好选择股票进行投资时，应当加强对股票基本信息的研究和长期的数据分析。由于目前中国证券市场还不够完善，资产重组的透明度不高，不排除存在一些股票依靠的是玩报表财技，如资产置换、变卖主业资产或者由大股东输送利润来扭亏的[7]，这些股票实际的生产经营并未得到改变，投资者在操作时要尽量谨慎。

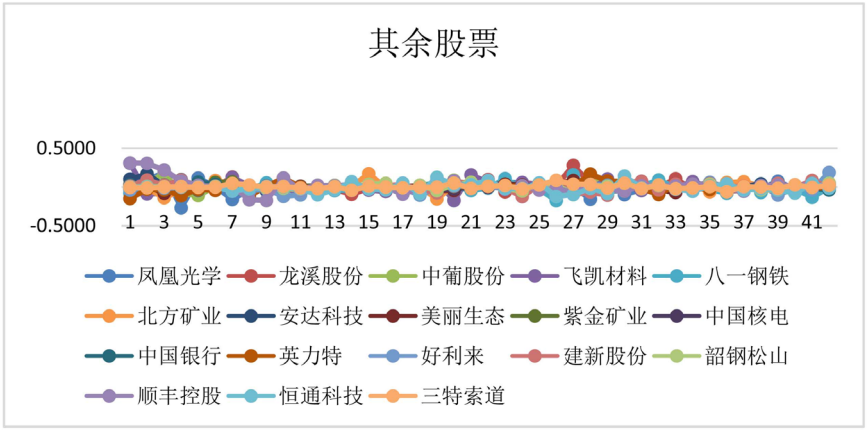

Figure 3. Line chart: stock yield of other stocks
图 3. 其余股票收益率波动折线图

6. 结束语

利用本文所述方法，可反映上市公司股票的盈利能力，有利于投资者缩小投资的范围，确定投资的价值，降低投资的风险，提高投资盈利的准确性，投资者可以重点关注有成长能力优势和盈利能力优势的稳定型和增长型公司。与此同时，在日常监管工作中，聚类分析模型也发挥了较大的积极作用，目前关于聚类分析在股票市场运用的文献浩如繁星，聚类方法也在不断得到改进，未来我们需要选择更加合适、更加完善的模型进行股票分析，进一步改进、完善聚类分析模型在股票市场的应用，稳定股票市场，促进实体经济的发展将成为主流的研究方向。股票市场的稳定繁荣是经济稳定繁荣的重要保障，更是人们生活水平提高的重要保障。

基金项目

杭州电子科技大学学生科研立项项目资助。

参考文献

[1] 李庆东. 聚类分析在股票分析中的应用[J]. 辽宁石油化工大学学报, 2005, 25(3): 94-96.
[2] 李慧. 聚类分析在股票投资分析中的应用[J]. 商, 2015(27): 199.
[3] 吴曼琪. 基于 K 均值聚类的 ST 股票分类研究及投资策略[J]. 中国城市经济, 2010(8X): 26-26.
[4] 洪月华. 蜂群 K-means 聚类算法改进研究[J]. 科技通报, 2016, 32(4): 170-173.
[5] 王强. 聚类分析模型在股票市场的应用[J]. 经济界, 2016(5): 101.
[6] 邓攀. 湖南省工业企业资本效率的实证研究[D]: [硕士学位论文]. 长沙: 中南大学, 2006.
[7] 王标. “乞丐帮主”的 ST 股操作全攻略[J]. 私人理财, 2005(4): 66-68.